A two factor linear model in bodyfat

Feiyun Yan, Srivats, Ziang Zeng

Background and data cleaning

Our group report focuses on building a two-factor linear regression model to describe the relevance between the bodyfat and other indexes of human body. Obesity and related health issues have become a global concern in recent years. However, measuring the body fat directly can be impractical or inconvenient. Understanding the factors that contribute to bodyfat can help us predict the index, reveal the factor relating to obesity and overall health, which can be helpful for public health, medical research, and individual well-being.

Our data contains 252 observations of 17 variables without missing value. From the age distribution we can see the data can not represent the whole population, the minimum age is 22. There are outliers based on body fat calculation. We found 4 rows doesn't follow Percentage of Body Fat (i.e. 100*B) = 495/D – 450 known as Siri formular considering the round-off error. We remove the negative or near-zero value of bodyfat after recovering it from the density.Moreover, we found that No.216 has highest body fat which is obviously abnormal. Sample 216 has higher body fat compared with Sample 39 whose body indexes are greater than his so we removed 216. Similarly, the outliers that violate the BMI formular 703 x weight (lbs) / [height (in)]^2 are removed because we don't know which one is the original data or backup. As for the redundant variable, we choose adiposity (BMI) other than weight and height and body fat other than density.

Final model

The final model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ whereβ_0 is the intercept, β_1 is the coefficient of waist hip ratio which is **abdomen/hip**, β_2 is the coefficient of adiposity(BMI).

$$Bodyfat = -65.06 + 70.47 \times Waist\ hip\ ratio + 0.7456 \times Adiposity$$

We assume the model follow the basic linear model assumption: independently normally distributed residual, no perfect multicollinearity, residuals are independent to response variable. We choose waist hip ratio because the ratio is commonly used in clinics to measure the obesity. Human body tends to put the fat tissue around the abdomen and organs to keep balance, and hip can show the skeleton size of the human body which doesn't store fat as much as abdomen.D.C. Chan et al. That's why we are interested in the relationship between the ratio and human body fat.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-65.0610590	5.1595907	-12.609732	<2e-16
waist_hip_ratio	70.4753299	7.1207179	9.897223	<2e-16
ADIPOSITY	0.7456411	0.1126268	6.620461	3.89e-10

For example, a man with 1 waist hip ratio and 25 in BMI is expected to have a body fat at 19.16%. The estimated parameters of waist hip ratio and adiposty are 70.475 and 0.746 which means one unity change in waist hip ratio can leads to 70.475 unit change in body fat, one unity change of adiposity can leads to 0.746 unit change of body fat.

Rationale and Model Diagnostics

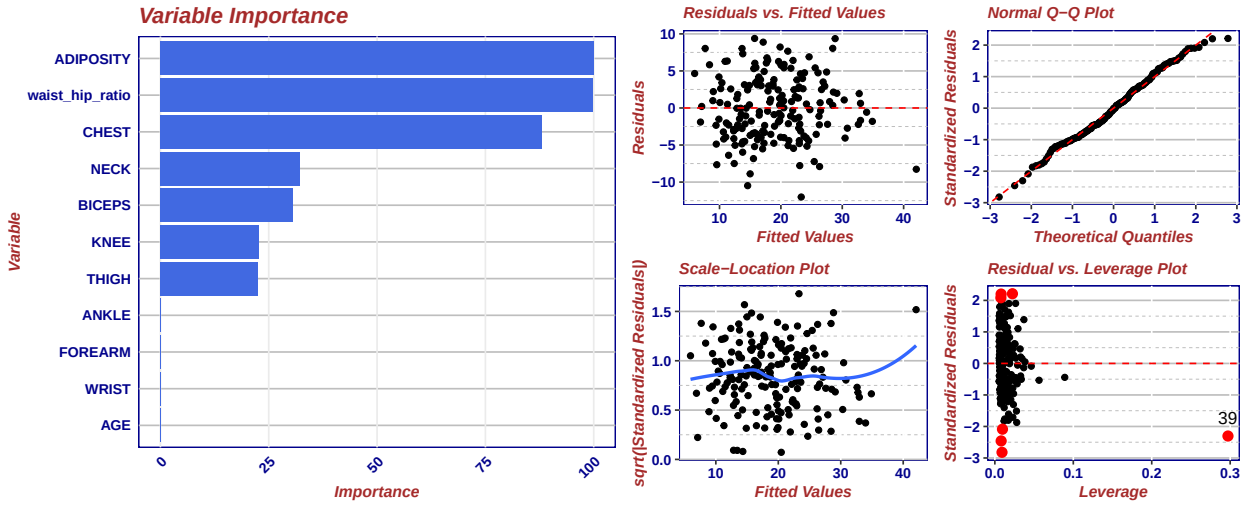
As shown before all the coefficients passed the t-test of significance under $p = 0.05$, and also the model passed the F-test with $p < 2e - 16$ by which we conclude that the model is reasonable. The R square in testing data is 0.6904 after bootstrapping which means 69.04% percent of the variance of body fat is explained by the predictors.

We choose our model because among all the linear models containing the ratio our model has the best r square in the training data which indicate a good accuracy, as for the testing data, the model with Thigh beats our model but it has only 0.606 R square in testing data which indicates over-fitting and lake of

robustness. We also tried full model random forest, they have worse performance in r square and robustness. PCA solves the multicllinearity but can not be easily interpreted.

Model	Adiposity	Neck	Thigh	Chest	Reg Tree
Training_rsq	0.691	0.627	0.710	0.667	*root node error*
Training_mse	18.03	21.75	16.89	19.42	*58.25*
Testing_rsq	0.655	0.598	0.606	0.593	0.4969
Testing_mse	20.21	23.51	23.20	23.80	30.13

Although the regression tree doesn't has good results as linear model, it also provides evidence in variable selection through the variable importance. Similarly the variable selection through BIC principle suggest waist_hip_ratio, adiposty, neck and thigh as predictors in the model has least BIC value. However, including all these variable leads to significant multicollinearity and weak robustness.



The plots above shows the verification of the assumptions mentioned before, in which the residuals is independent to fitted values because there is no obvious pattern and the white's test accept H_0 of Homoskedasticity with $p = 0.8359$. Also the residuals are normally distributed which pass the Shapiro normality test with $p = 0.2937$(can not reject H_0). To test multicollinearity, we use VIF which shows a vif value $1.778 < 5$ meaning there is no obvious multicollinearity. At last, the residuals vs leverage plot shows there exists a outlier No.39 who has largest weight in the data. Removing this outlier, we saw a increase in the r square to 0.6977 in training data.

Conclusion

In conclusion, our two-factor linear model shows the relationship between the body fat and waist hip ration as well as BMI, one can use our model to measure their body fat with index can be conveniently measured. Due to the simplicity of our model, one can easily understand and apply it to new data.

However, there are still weakness of our model. Firstly, our r square in the testing data is not good enough, which implies there are still other effects of body fat that our model hasn't include, for example the nonlinear effects. Secondly, we don't consider about the interaction effect of our predictor in the model which could be a future work.

Contribution: Srivats makes the PPT and the presentation, Ziang Zeng completes the summary and the modeling code. Feiyun Yan make shiny app and provide key inspiration of model, we do the pre-modeling and EDA together.

References

Chan, D.C., G.F. Watts, P.H.R. Barrett, and V. Burke. "Waist Circumference, Waist-to-Hip Ratio and Body Mass Index as Predictors of Adipose Tissue Compartments in Men." *QJM: An International Journal of Medicine*, vol. 96, no. 6, June 2003, pp. 441-447. https://doi.org/10.1093/qjmed/hcg069.